Assessing the Readiness of Higher Education Institution in Malaysia to Accept Generation Alpha using Predictive Analytics

Tan Lik Wei, Mohd Norshahriel Abd Rani, and Nabilah Filzah Mohd Radzuan

INTI International University, Malaysia, {i17012619@student.newinti.edu.my, mnorshahriel.arani@newinti.edu.my, nabilah.radzuan@newinti.edu.my}

ABSTRACT

The impact of technology in education is getting more common. In the UK, most of the children start to learn coding skills from the age of 5. Generation Alpha will be very different from traditional college students. Technologies will be largely driven in education and educators need to learn how to adapt to it. The traditional method of teaching and learning might not be effective and efficient for Generation Alpha. Institutional culture needs to be changed to prepare the arrival of Generation Alpha students. In a technology-driven period, students need to learn problem solving skills to help themselves how to think not what to think, and collaboration skills to collaborate with peers around the world. What will the Generation Alpha students behave in higher education? How to define that an institution is ready to accept Generation Alpha? These questions can be answered by finding the unique pattern of generation z using predictive analytics. This research is focus on develop a dashboard system in assist decision making for the higher education institution. The dashboard system will allow Higher Education Institution (HEI) to capture and analyze useful insights and improve decision making from the student data. K-nearest Neighbours (KNN), Support Vector Machine (SVM), and XGBoost are data mining techniques are implemented in order to develop the prediction analytics model. The testing will be conduct for analysis and evaluation of the system.

Keywords: generation alpha, higher education institution, supervised learning, data mining, KNN, SVM, XGBoost.

I INTRODUCTION

The impact of technology in education is getting more common. Generation Alpha will be very different from traditional college students. Technologies will be largely driven in education and educators need to learn how to adapt to it. The traditional method of teaching and learning might not be effective and efficient for Generation Alpha. Institutional culture needs to be changed to prepare the arrival of Generation Alpha students. The current generation of the university or college students is called Generation Z who are born after the year 2000 (Pikhart and Klímová 2020). This group of children is born since 2010, which is the year that Apple unveils the first iPad (Apple 2017). They are the first generation that grows up in a digital world with exposure to electronic devices.

Education has been transformed from passive and reactive to interactive and aggressive (Raja and Nagasubramani 2018). Apple has been an aspiring kid coder who start as young as 6 years old (Apple 2018) which is unlikely to observe in previous generations (Romero Jr 2017). Generation Alpha is a generation that chooses technology over a human connection which loses human connection, but they will be the most educated generation (Romero Jr 2017).

The technology has been integrated into their lives seamlessly (Hughes 2020). Judy Raper who is an engineering school founder told Nature Index "We will expose students to problem-solving in industry, which is often way ahead of us," (Dall et al. 2018). The Generation Alpha will be not the same as the traditional college students that we are seeing now (Romero Jr 2017). There are 41 out of 144 papers related to inappropriate curriculum and teaching strategy issue which indicates the importance of quality in learning (Quadir, Chen, and Isaias 2020).

Therefore, the scope of this research will focus on predictive data mining model for Generation Alpha students would be developed to analyze and extract previously unknown patterns. A dashboard system which is a web-based platform will be created for the Higher Education Institution (HEI) to analyze key metrics, visualize insights, and identify bottlenecks of the current solution.

The dashboard system is targeted for the Higher Education Institution (HEI) which includes most public and private institutions in Malaysia. The users should be able to open the website the obtain useful insights that will help the institutions to facilitate decision making.

II BACKGROUND STUDY

This reviews the existing body of literature in the context of behavioral characteristics of Generation Alpha and the future of Higher Education Institution (HEI). The various predictive techniques are assessed and explained in detail.

A. Generation Alpha

Generation Alpha is the children who born years 2010 onwards (Velički and Velički 2015). Generation Alpha is far from being a household name to describe the new wave of world inhabitants, but it is one of the terms being used to describe those being born at the cross-over of Generation Z and new age.

B. Future of Higher Education Institution (HEI)

The future of Higher Education Institution (HEI) in Malaysia is uncertainty. 38% of 18-year-old applicants in the UK received at least one unconditional offer in the year 2019 (Johnson 2020). An unconditional offer is an offer that does not rely on the results of previous studies.

C. Supervised Learning

A subcategory of machine learning and artificial intelligence is supervised learning, also known as supervised machine learning (Education 2020). The other subcategories of machine learning are unsupervised learning and reinforcement learning. There are several types of classification models include Logistic Regression, K-nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest and Gradient Boosting. The author has selected K-nearest Neighbours (KNN), Support Vector Machine (SVM), and XGBoost to develop the prediction model.

For the first step of the KNN model is to select the number of K of neighbours. When K=1, then the algorithm is known as the nearest neighbour algorithm which is the simplest case (Datacamp 2018). Support Vector Machine (SVM) was initially constructed in the 1960s and then were refined again in 1990s. It is different compared with other machine learning algorithms. SVM is a model that trying to find the best decision boundary which helps to separate a space into two classes.

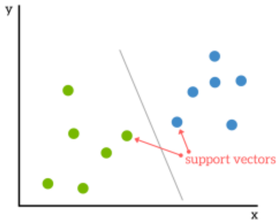


Figure 1. Support Vector Machine (SVM)

XGBoost is focused on execution speed and model performance. It is fast compared to other models. XGBoost is widely used in structured datasets on classification and regression problems.

III RESEARCH METHODOLOGY

Knowledge Discovery in Databases (KDDs) is the programmed extraction of implicit and useful insights from the huge amount of data collections (Klösgen and Zytkow 2002). The KDD process model is selected for the purpose system because it is complete and more precise compared with CRISP-DM and SEMMA which are mostly company-oriented (Shafique and Qaiser 2014) as shown in Figure 2.

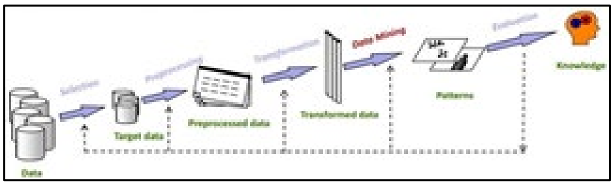


Figure 2. Steps in KDD (Fayyad, et al., 1996)

IV SYSTEM DESIGN

The purpose is to provide specifications of the system functions or features for the end-user. Requirements capture is important to understand the requirements of the client.

A. Fact-Finding Techniques

Three interviews have been conducted through a video-conferencing application from Microsoft Teams. The interviewees can answer the questions freely because of its open-ended questions. One lecturer and two students are selected for the interview session.

The world is preparing for Generation Alpha, Higher Education Institution (HEI) in Malaysia has to prepare for Generation Alpha. Generation Alpha is unlike the previous generation which has grown in various emerging technology. Higher Education Institution (HEI) in Malaysia has implemented blended learning for Generation Z students, they may have planned something for the Generation Alpha students. They may have a new module to teach the Generation Alpha.

It can be possible to happen that online class will be replaced by physical classes in the future because the difference between online classes and physical classes is the medium of learning, but the knowledge can be shared to the students in the same method.

It can be possible to happen because the difference between an online class and physical class is the medium of learning, but the knowledge can be shared to the students in the same method. Students do not want to open their camera and the educators could not observe the students.

V IMPLEMENTATION

A. Exploratory Data Analysis (EDA)

The following bar chart in figure 3 shows the number of exercises in each learning stage. The elementary learning stage has 784 exercises which own the highest quantity, junior learning stage has 543 exercises, and senior learning stage has 3 exercises which own the lowest quantity.

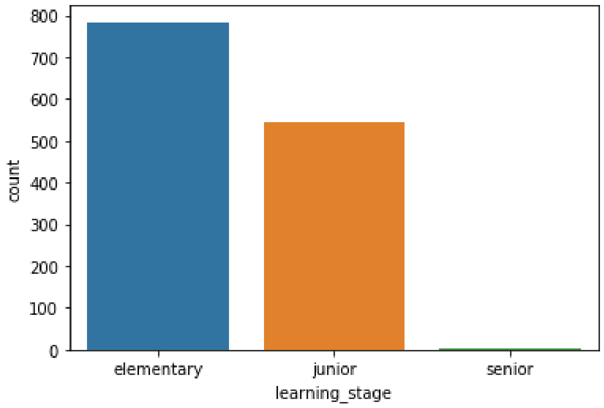


Figure 3. Distribution of Learning Stages

The number of exercises in various difficulties shown in figure 4. The easy difficulty has 784 exercises which own the highest quantity, junior learning stage has 543 exercises, and senior learning stage has 3 exercises which own the lowest quantity.

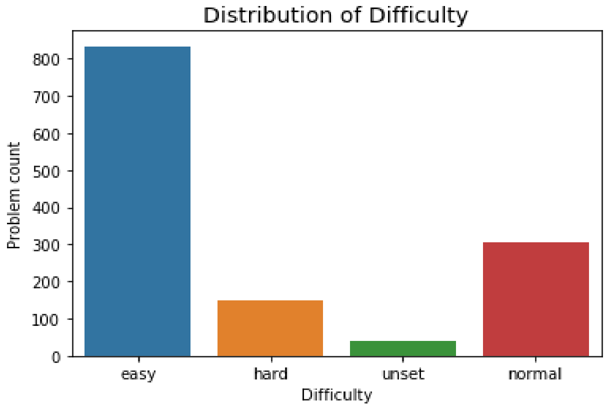


Figure 4. Distribution of Difficulties

The count plot of the target variable which is the level column is shown in figure 5. This could know how the level distributions are among the courses. It helps to find out any improvement required in the course of Junyi Academy.

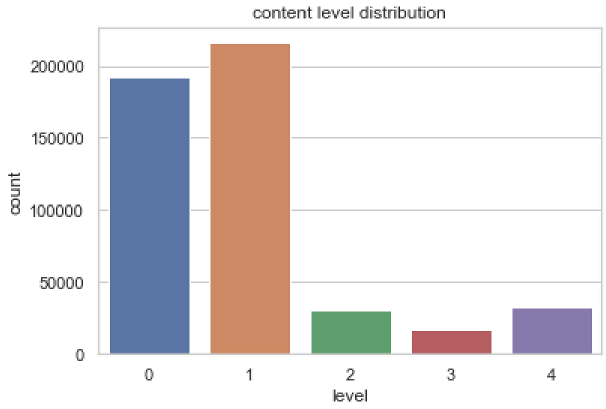


Figure 5. Distribution of Content Levels

From the distribution plot in figure 6, it indicates that the highest number of problems attempts in the exercise is 5 or 6 problems. Because of the proficiency mechanism they might want to upgrade their level to become level 1 and continue with the next exercise.

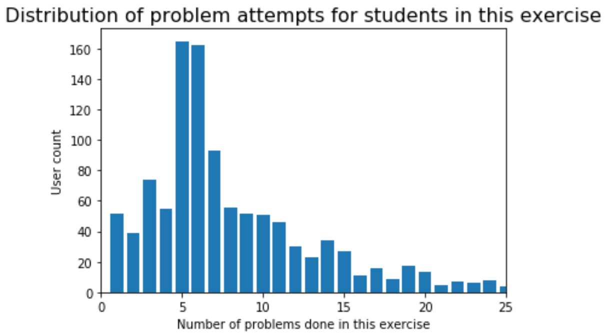


Figure 6. Distribution of Problem Attempts for Students in Exercise

B. Resampling

Resampling is a commonly used technique for interacting with extremely unbalanced datasets (Vidhya 2020). If the dataset is not balanced, the outcomes of the prediction model might not be accurate to deploy into a production model. The target variable which is the level attribute is visualized with the following plot. The plot figure 7 shows that there is a large gap among the datasets. The number of students who are having the level of 0 and 1 is more than other levels.

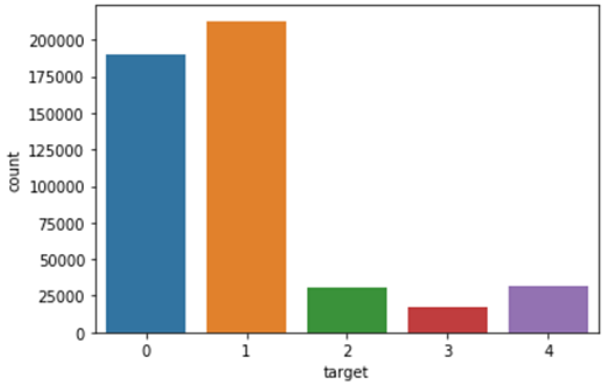


Figure 7. Target Variable Distribution Before Resampling

C. Feature Selection

The variable has to be reviewed and dropped if the variable has a very low variation because the model is not going to learn anything with it. If the variance is low or close to zero, then a feature is approximately constant and will not improve the performance of the model. In that case, the variable who results in zero variation should be removed.

$$VAR(x) = \frac{1}{\eta}\sum_{\iota=1}^{\eta} (x_{\iota} - \mu)^2 \qquad (1)$$

D. Feature Importance

Feature importance applies to strategies that assign a score to the input function based on how beneficial a target variable can be expected (Brownlee 2020). This technique could provide insights into the dataset and model by selecting the most appropriate attributes for predictive modelling.

Table 1. Feature Permutation Importance

Weight	Feature
0.5884 ± 0.0041	is_upgrade
0.3745 ± 0.0031	problem_number
0.1002 ± 0.0014	total_attempt_cnt
0.0999 ± 0.0006	is_hint_used
0.0966 ± 0.0011	points
0.0896 ± 0.0018	used_hint_cnt
0.0752 ± 0.0007	total_sec_taken
0.0748 ± 0.0011	badges_cnt

The result of feature importance in table 1 shows that the 8 variables which are most important in the dataset. These variables would be stored for data modelling. After the unused attributes have been dropped, the author split the data to train and test set. 20% of the data is used to create the test data and 80% to create the train data.

E. *Principal Component Analysis (PCA)*

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

After applying PCA, the training set has reduced the number of attributes to 7. An attribute which does not do any impact to the target variable has been removed from the data frame.

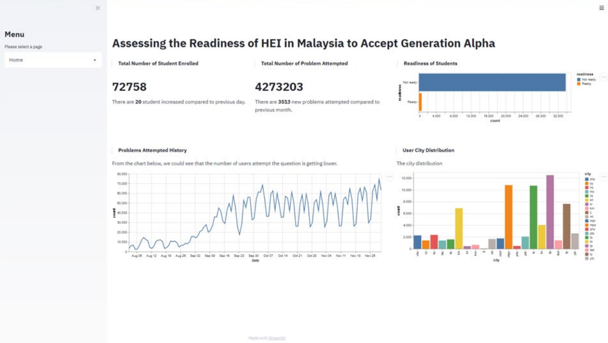
F. System Dashboard

The home page is a single page dashboard which consists of the highlights of assessing the readiness of HEI in Malaysia to Accept Generation Alpha and obtain the number of students is performing well in the courses at the top-right of the page.

The total number of students enrolled and problems which have been attempted in the system are listed. The dashboard also provides information about the increased number of students or problems to allow users to acquire information about the system performs compared with the previous month.

The line chart indicates the timeline of the problem attempted by the students. From figure 8, the author could determine that the number of problems attempted is increased by each month. Seasonality can be found in the chart.

The user city distribution shows the number of user account based on different cities in Taiwan. From the bar chart, there are 20 cities in Taiwan which can be differentiated by the colours.



Assessing the Readiness of HEI in Malaysia to Accept Generation Alpha

Total Number of Student Enrolled: 72758

There are **20** student increased compared to previous day.

Total Number of Problem Attempted: 4273203

There are **3513** new problems attempted compared to previous month.

Problems Attempted History

From the chart below, we could see that the number of users attempt the question is getting lower.

User City Distribution

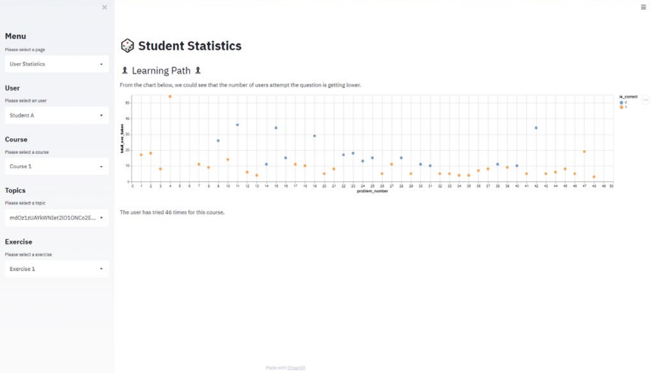
The city distribution

Figure 8. Home Page of the Dashboard

The student statistics page consists of a detailed breakdown of the learning path of the course. It provides a drill-down view to have a more detailed insight from the dataset.

The scatter plot in figure 9 shows how the student performs courses. It uses two colours to differentiate the correctness of the exercises. From the x-axis, it is the time taken for each exercise, the higher the x value means that the user has taken more time at the exercise.

The users can select the users, courses, topics, and exercises that they wanted to in-depth analyse the student behaviour. The system also indicates the number of times that the students have attempted in the course.



Student Statistics

Learning Path

From the chart below, we could see that the number of users attempt the question is getting lower.

The user has tried 46 times for this course.

Menu: User Statistics

User: Student A

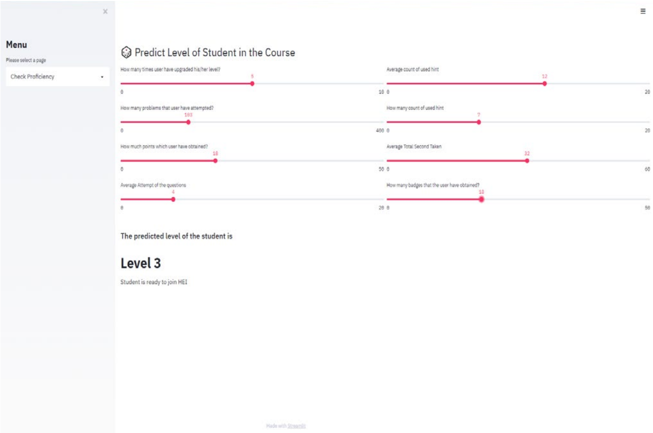
Course: Course 1

Topics: mdOz1zUAYkWNIet2lO1ONCo2E...

Exercise: Exercise 1

Figure 9. Student Statistics Page of the Dashboard

The system can predict the level of the student based on the factors which are important. These factors or features are identified in the feature selection phase. There is a total of 8 important features. The system also provides information that the student is ready to join HEI or not based on the student's performance; figure 10.



Predict Level of Student in the Course

How many times user have upgraded his/her level? 5 (0–10)

Average count of used hint 12 (0–20)

How many problems that user have attempted? 183 (0–400)

How many count of used hint 7 (0–20)

How much points which user have obtained? 18 (0–50)

Average Total Second Taken 32 (0–60)

Average Attempt of the questions 4 (0–20)

How many badges that the user have obtained? 18 (0–50)

The predicted level of the student is

Level 3

Student is ready to join HEI

Figure 10. Student Level Prediction Page of the Dashboard

VI EVALUATION

Evaluating a model is an important phase of building an effective machine learning model (Srivastava 2019). There are several types of evaluation metrics, but not every type is suitable for all kinds of machine learning models. Precision is defined among the examples which are projected to belong to a certain class as the fraction of relevant examples (Jordan 2017). This metric is used where the correct prediction is important.

$$P = \frac{T_p}{T_p + F_p} \tag{2}$$

Recall is described as the fraction of the examples predicted to belong to a class in relation to all the cases that really belong to the class (Jordan 2017)

$$R = \frac{T_p}{T_p + F_n} \tag{3}$$

These quantities are also related to the (F1) score, which is defined as the harmonic mean of precision and recall (scikit-learn 2020).

$$F1 = 2\frac{P \times R}{P + R} \tag{4}$$

The precision and recall cannot be maximized in both ways because increasing the precision score would decrease the recall score, otherwise increasing the recall score would decrease the precision score.

It would be better to maximize the precision because the model is required to predict correctly when the student is detected as ready to join HEI. It is not ideal to have a model that predicted a proficient student as not ready to join HEI.

K-NN has a weighted average precision score of 0.785, recall score of 0.787 and f1-score of 0.786 are shown in Table 2.

Table 2. Classification Metrics for K-NN Model

Precision	Recall	F1-Score
97.9%	99.7%	98.8%
85.1%	94.6%	89.6%
68.6%	76.7%	72.4%
62.5%	57.7%	60.0%
79.3%	65.9%	72.0%

SVM has a weighted average precision score of 0.779, recall score of 0.777 and f1-score of 0.761 are shown in Table 3.

Table 3. Classification Metrics for SVM Model.

Precision	Recall	F1-Score
98.9%	100%	99.5%
76.2%	99.9%	86.5%
66.1%	83.7%	73.9%
73.7%	37.5%	49.7%
75.3%	68.1%	71.5%

XGBoost has a weighted average precision score of 0.796, recall score of 0.793 and f1-score of 0.791 are shown in Table 4. All the metrics for this model had performed very well. The precision score is the highest which means that the performance of the model is good compared with the previous models.

Table 4. Classification Metrics for XGBoost Model

Precision	Recall	F1-Score
98.3%	99.6%	98.9%
86.9%	94.5%	90.5%
68.7%	79.0%	73.5%
62.0%	61.6%	61.8%
82.9%	62.8%	71.5%

The analysis has been done on the student part. As there are different kinds of factors or features which might affect the level of proficiency of a student. The users could obtain information about the students' performance in school. They can have taken some actions to tackle the problems of the students which might affect their proficiency in the courses.

VII EVALUATION RESEARCH LIMITATION

Some potential limitations could be addressed in future research. Generation Alpha does not exist much data yet and makes it hard to do proper research which is related to Generation Alpha. The result concludes that there are not many research papers or journals that can be obtained. Defining the characteristics of Generation Alpha is difficult compared to previous generations. Due to the limited access of Generation Alpha data, the author would like to acquire the data from the current generation which is Generation Z. These data can be useful to analyze the insights and predict the patterns for Generation Alpha.

VIII CONCLUSION

The author has research on the supervised learning models which would be used to predict or classify the student's performance. The author has also researched about the tools and techniques that would be applied in the implementation phase. A dashboard has been developed by using Streamlit. The classification metrics have been listed and compared among the machine learning models. The dashboard system is also being evaluated.

REFERENCES

Apple. 2017. "Apple Launches IPad, Apple Newsroom." Retrieved June 8, 2020 (https://www.apple.com/au/newsroom/2010/01/27Apple-Launches-iPad/).

Apple. 2018. "Apple Brings Coding Education to More Students for Computer Science Education Week." Retrieved June 8, 2020 (https://www.apple.com/au/newsroom/2018/11/apple-brings-coding-education-to-more-students-for-computer-science-education-week/).

Dall, Writers Imogen, Dof Dickinson, Rodney Payne, and Sean Tierney. 2018. *Transforming Education: Empowering the Students of Today to Create the World of Tomorrow*.

Datacamp. 2018. "KNN Classification Using Scikit-Learn." Retrieved November 13, 2020 (https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn).

Education, IBM Cloud. 2020. "Supervised Learning."

Hughes, Joanna. 2020. "Getting to Know Generation Alpha: 10 Takeaways for Higher Ed." Retrieved June 8, 2020 (https://www.keystoneacademic.com/news/getting-to-know-generation-alpha-10-takeaways-for-higher-ed).

Johnson, Paul. 2020. "Automation and Employment: How Universities Must Respond to Industry 4.0." *The Drum.* Retrieved July 6, 2020 (https://www.thedrum.com/opinion/2020/01/30/automation-and-employment-how-universities-must-respond-industry-40).

Jordan, Jeremy. 2017. "Evaluating a Machine Learning Model." Retrieved November 15, 2020 (https://www.jeremyjordan.me/evaluating-a-machine-learning-model/#:~:text=The three main metrics used,the number of total predictions.).

Klösgen, Willi, and Jan M. Zytkow, eds. 2002. *Handbook of Data Mining and Knowledge Discovery.* USA: Oxford University Press, Inc.

Pikhart, Marcel, and Blanka Klímová. 2020. "ELearning 4.0 as a Sustainability Strategy for Generation Z Language Learners: Applied Linguistics of Second Language Acquisition in Younger Adults." *Societies* 10(2):38.

Quadir, Benazir, Nian Shing Chen, and Pedro Isaias. 2020. "Analyzing the Educational Goals, Problems and Techniques Used in Educational Big Data Research from 2010 to 2018." *Interactive Learning Environments* 0(0):1–17.

Raja, R., and P. C. Nagasubramani. 2018. "Impact of Modern Technology in Education." *Journal of Applied and Advanced Research* 3(S1):33–35.

Romero Jr, Aldemaro. 2017. "Colleges Need to Prepare for Generation Alpha." *The Edwardsville Intelligencer* 3.

scikit-learn. 2020. "Precision-Recall." Retrieved October 26, 2020 (https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html).

Shafique, Umair, and Haseeb Qaiser. 2014. "A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)." *International Journal of Innovation and Scientific Research* 12:2351–8014.

Srivastava, Tavish. 2019. "11 Important Model Evaluation Metrics for Machine Learning Everyone Should Know." *Analytics Vidhya.* Retrieved November 13, 2020 (https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/).

Vidhya, Analytics. 2020. "10 Techniques to Deal with Imbalanced Classes in Machine Learning." Retrieved November 12, 2020 (https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/).